Janusz S. Bień
Uniwersytet Warszawski

Podstawowe elementy tekstów elektronicznych

Wstęp

Znaczenie języka pisanego z punktu widzenia badań lingwistycznych zmieniało się z czasem – wielu językoznawców uważało, że jedynym przedmiotem godnym zainteresowania jest język mówiony[1]. Inne stanowisko zajmuje Marek Świdziński – należy on bowiem do tych badaczy, którzy w sposób jawny koncentrowali się na języku pisanym. Pokazują to choćby tytuły jego artykułów: *Klasyfikacja prostych grup syntaktycznych we współczesnej polszczyźnie pisanej* (1979), *Formalna definicja równorzędnej grupy nominalnej we współczesnej polszczyźnie pisanej* (1981a, współautor Stanisław Szpakowicz), *Szkic koncepcji ogólnego słownika podstawowego współczesnej polszczyzny pisanej* (1982, współautorzy: Zygmunt Saloni, Stanisław Szpakowicz), *Zarys klasyfikacji schematów zdaniowych we współczesnej polszczyźnie pisanej* (1981b, współautor Stanisław Szpakowicz), *Tzw. wypowiedzenie złożone w gramatyce formalnej współczesnej polszczyzny pisanej* (1984), *Elipsa w gramatyce formalnej współczesnej polszczyzny pisanej* (1986). Takie podejście do języka najpełniejszy wyraz znajduje w książce *Gramatyka formalna języka polskiego*, w której czytamy między innymi, por. Świdziński (1992, 20):

> Interpunkcję uważam za zjawisko czysto składniowe, a znaki przestankowe (przynajmniej niektóre) traktuję jako składniki bezpośrednie odpowiednich konstrukcji składniowych.

Należy jednak pamiętać, że wyrażenie „język pisany” jest wieloznaczne, może odnosić się zarówno do tekstów pisanych ręcznie, jak i drukowanych , a także do tekstów wprowadzonych bezpośrednio z klawiatury do komputera lub innego urządzenia (np. telefonu komórkowego). O tych tekstach mówi się, że są one „uro-

[1] Współczesne językoznawstwo ujmuje język w ramach ogólniejszego zjawiska mowy. Mowa to wszelki fakt dźwiękowego porozumiewania się ludzi, por. STJ (1970, 271).

dzone cyfrowo" (ang. *digitally born*). Te różne warianty tekstów pisanych różnią się istotnie, należy je więc traktować w zróżnicowany sposób.

O ile teksty pisane ręcznie pod wieloma względami przypominają teksty mówione (Bień 1991, 9), o tyle typowe teksty drukowane za pomocą tzw. składu gorącego z tekstami zapisanymi w komputerze łączy jedna wspólna cecha – dają się w nich jednoznacznie wyróżnić ich podstawowe niepodzielne jednostki. W wypadku tekstu drukowanego jednostki te stanowią odwzorowanie fizycznych obiektów służących do zbudowania matrycy drukarskiej, nazywanych czcionkami – termin ten ma podobno około 200 lat. Żegota Wywiałkowski (1881, 21) pisał w swoim słowniczku pod hasłem *krotły*:

> Jest to najdawniejszy i może najpierwszy wyraz polski, a który w użyciu zaniedbany został, i zastąpiono go wyrażeniami: Typy, Charaktery, Pismo, aż około 1840 r. poczęto używać wyrazu Czcionki. O ile mogłem dojść, nasz myśliciel Bronisław Trentowski, pierwszy użył tego słowa pisząc Trzcionki, które to wyrażenie otrzymało obywatelstwo w Słowniku polskim [zapewne chodzi o słownik Lindego – JSB], ale ze zmienioną pisownią: Czcionki.

Warto zwrócić uwagę na występujące w cytacie słowo *charakter*, które w innej pisowni występuje w tytule znanego traktatu Januszowskiego (1594). Pochodzi ono z języka greckiego, w którym – według Wikipedii (Character 2011) – oznaczało początkowo między innymi wygrawerowaną pieczęć, wtórnie – jej odbicie, a następnie znaki pisma, zwłaszcza obcego. Za pośrednictwem łaciny i języka francuskiego słowo to weszło do języka angielskiego, w którym w interesującym nas znaczeniu pojawia się (znów według Wikipedii) po raz pierwszy w XV wieku. Z czasem zaczęto je odnosić do wszelkich znaków piśmiennych, nie tylko obcych, lecz także tych tworzonych mechanicznie za pomocą maszyn do pisania czy dalekopisów. Po pojawieniu się komputerów termin ten zaczął być używany w znaczeniu podstawowego elementu wszelkich napisów, początkowo reprezentowanego przez 6 lub 7 bitów, później przez pojedyncze 8-bitowe bajty, a obecnie teoretycznie nawet przez 4 bajty, czyli 32 bity.

W języku polskim angielskiemu terminowi *character* odpowiada termin *znak*, jednak zbyt ogólne skojarzenia związane z tym słowem są w niektórych kontekstach mylące.

Ponieważ rewolucja informatyczna objęła również nauki humanistyczne, obecnie przedmiotem badań lingwistycznych są również teksty elektroniczne (także teksty powstałe w wyniku dygitalizacji tekstów fizycznych).

1. Grafemy

Pojęcie grafemu zostało wprowadzone w 1915 r. przez Jana Niecisława Baudouina de Courtenay (1983, 84): „Istniejące stale w naszej psychice wyobrażenie niepodzielnej litery nazywamy *grafemą*" – jak widać, początkowo termin ten był

rodzaju żeńskiego. W praktyce jest on stosowany rzadko, pomija go wiele słowników i encyklopedii.

Ciekawym przykładem wykorzystania pojęcia *grafemu* są prace Tomasza Lisowskiego, w szczególności książka (Lisowski 2001) i artykuł (Lisowski 2004). Inspiracją dla autora były z jednej strony prace Adama Heinza na temat pojęcia *wariantu językowego*, z drugiej zaś prace Piotra Ruszkiewicza dotyczące grafemów w języku angielskim. Lisowski (2001, 16) definiuje *grafem* jako 'dyspozycję do graficznego substytuowania fonemu', która jest realizowana w tekście przez *alografy*. Wyróżnia się *alografy tekstowe* i *fakultatywne* oraz *prymarne* i *sekundarne*. Wprowadzony przez autora aparat pojęciowy wydaje się jednak zbyt prosty i niewygodny (co nie umniejsza wartości i użyteczności tych publikacji).

Alografy fakultatywne w pracach Lisowskiego nie odgrywają istotnej roli, pojęcia te zostały chyba wprowadzone przede wszystkim po to, aby jakoś określić te aspekty tekstów, które znajdują się poza zakresem zainteresowań autora. Ich definicje budzą jednak wątpliwości, które warto odnotować. *Alograf fakultatywny prymarny* to 'wariant stylistyczny grafemu, którego zasięg użycia może być arbitralnie regulowany (chodzi o takie rozstrzygnięcia jak stosowanie wielkich i małych liter, spacji)' (Lisowski 2001, 17). *Alograf fakultatywny sekundarny* to 'wariant stylistyczny grafemu, którego zasięg użycia jest regulowany indywidualnymi upodobaniami estetycznymi (chodzi tu o krój czcionki, np. fraktura, szwabacha, antykwa itd., lub też o indywidualny charakter pisma odręcznego)'. Wątpliwości budzi nazywanie tych wariantów stylistycznymi, a także stwierdzenia, że stosowanie wielkich i małych liter jest regulowane arbitralnie (przez kogo?) oraz że wybór fontu to wyłącznie kwestia estetyczna – gdyby niniejszy tekst został wydrukowany frakturą, to zostałby odrzucony przez wydawcę wcale nie z powodów estetycznych... Zasadnicza jednak słabość tych definicji polega na tym, że alografy fakultatywne traktuje się równorzędnie z alografami tekstowymi jako reprezentacje grafemów, podczas gdy w rzeczywistości opisują one dodatkowe własności elementów tekstowych, czyli – w terminologii autora – alografów tekstowych.

Alograf tekstowy u Lisowskiego (2001, 17) to przede wszystkim litera alfabetu, niekiedy na gruncie polskim wzbogacona o znak diakrytyczny (kreska, kropka, haczyk, ogonek) bądź o literę-diakryt (tzw. dwuznaki ſſ lub trójznaki ſſh), ale też kombinacja litery (liter) z kontekstem graficznym (spacja, alografy innych grafemów).

Ostatni człon tego wyliczenia to, jak się wydaje, mylący skrót myślowy – kontekst wpływa na wybór tzw. alografu pozycyjnego, ale nie jest jego składnikiem. Jednak autorowi, omawiającemu szczegółowo realizacje grafemów w wybranych tekstach szesnastowiecznych, wygodnie jest w tabelarycznych zestawieniach traktować równorzędnie właściwe alografy i reguły kontekstowe.

Alograf tekstowy prymarny to „najczęstszy tekstowy wariant grafemu [...]" (Lisowski 2001, 16); jeśli jego frekwencja jest równa co najmniej 95%, to nazywany

jest on *fonografem*. W moim odczuciu termin ten jest zdecydowanie niezręczny, a samo pojęcie chyba zbędne. Arbitralność tej definicji szczególnie razi w kontekście stwierdzenia: „Ortografię tekstu tworzą więc fonografy, fonografy pozycyjne, alografy fakultatywne" (Lisowski 2001, 18).

O ile prace posługujące się jawnie pojęciem grafemu są rzadkie, o tyle jego niejawne zastosowania są znacznie szersze. Jak pokazałem w swojej książce (Bień 1991), interpretacja niektórych tabel fleksyjnych wprowadzonych do polskiej leksykografii przez Jana Tokarskiego jest możliwa tylko przy wykorzystaniu tego pojęcia. Jest ono również bardzo przydatne do formułowania współczesnych reguł ortograficznych, w szczególności zasad dzielenia wyrazów. Z tego względu postulowałem wprowadzenie pojęcia wyrazu *grafemicznego* i *grafemicznego poziomu reprezentacji tekstów*, ale nie było dotąd motywacji, aby te postulaty zrealizować w konkretnych programach przetwarzania języka naturalnego.

2. Unicode

W komputerach *znak* (ang. *character*) jest reprezentowany przez pewien zestaw bitów. Każdy taki zestaw bitów może być interpretowany jako liczba, którą nazywam *współrzędną kodową znaku* (ang. *code point*). Konwencja określająca interpretację współrzędnych kodowych to *kodowy zestaw znaków* (ang. *coded character set*).

Unicode to uniwersalny kodowy kod znaków, obejmujący docelowo wszystkie języki żywe i martwe. Jego koncepcja powstała pod koniec lat osiemdziesiątych ubiegłego wieku, a aktualna wersja to Unicode 6.0.0. Wersja ta została ogłoszona 11 października 2010 r., zawiera ona ponad 100 tysięcy znaków.

W pewnym stopniu znaki Unicode'u można uważać za wirtualne czcionki drukarskie, ale od czcionek różni je przede wszystkim to, że nie posiadają one konkretnego kształtu – kształty, nazywane *glifami*, są przypisane konkretnym współrzędnym kodowym dopiero w konkretnych fontach, np. litera *a* jest tym samym znakiem niezależnie od kroju i wielkości. Jednocześnie ze względów historycznych znakami są m.in. tzw. ligatury techniczne, np. *fi* – w technologii składu gorącego te dwie litery odlewano na jednym słupku z powodu „przewieszki", czyli nachodzenia oczka litery *f* na słupek litery *i*; przy okazji ze względów estetycznych litera *i* jako człon ligatury traciła swoją kropkę. Swoją drogą, umieszczanie na jednym słupku często stosowanych par znaków przyspieszało skład.

Również z czasów składu ręcznego wywodzi się konwencja, że często używane znaki z diakrytami stanowią samodzielne czcionki, ale oprócz tego do dyspozycji zecera są również „akcenty latające" dodawane w razie potrzeby do czcionek literowych w celu uzyskania rzadziej stosowanych liter obcych. W standardzie Unicode mamy również *znaki kompozytowe* (*kompozyty*, ang. *composite*

characters) oraz *znaki dostawne*[2], które tworzą całość razem z poprzedzającym je znakiem bazowym (ang. *base character*).

Konsekwencją tych własności Unicode'u jest to, że niektóre znaki i napisy mogą być reprezentowane na kilka sposobów. Litera *ń* może być zapisana jako kompozyt albo jako sekwencja dwóch znaków: litery *n* i akcentu akutowego; między tymi reprezentacjami zachodzi tzw. *równoważność pełna*, czyli *kanoniczna* (ang. *canonical equivalence*). W wypadku ligatury *fi* oraz sekwencji liter *f* i *i* mamy do czynienia tylko z równoważnością częściową nazywaną *równoważnością dostosowawczą* (ang. *compatibility equivalence*) – w niektórych kontekstach zastąpienie tej sekwencji liter przez ligaturę nie jest właściwe; równoważność dostosowawcza nie ma jednorodnego charakteru, niemal każdy wypadek jest nieco inny – tutaj warto wspomnieć, że relacja ta łączy w szczególności długie i krótkie *s*. Na potrzeby porównywania napisów i przeszukiwania tekstów niezbędne jest więc stosowanie normalizacji zapisu, którą można realizować, preferując kompozyty lub ich unikając, stosując przy tym równoważność kanoniczną lub dostosowawczą. Daje to w rezultacie 4 standardowe metody normalizacji.

W standardzie Unicode termin *grafem* występuje w bardzo ograniczonym zakresie. W słowniku terminów podane są dwie definicje, obie mało konkretne. Jedna z nich stwierdza, że dla konkretnego systemu pisma grafem to minimalna jednostka dystynktywna; definicja taka nie przesądza jednak, co jest minimalną jednostką – tylko litery czy również znaki diakrytyczne. Druga definicja podaje, że grafem to „to, co użytkownik uważa za znak (ang. *character*)", jest więc jeszcze mniej konkretna. Charakterystyczną cechą obu definicji jest ich względność. Ponieważ Unicode koncentruje się na opisie własności uniwersalnych, niezależnych od konkretnego języka naturalnego, w zasadzie nie ma potrzeby ani możliwości wykorzystania na jego potrzeby pojęcia grafemu, funkcjonuje natomiast w nim pokrewne pojęcie *zbitki grafemicznej* (ang. *grapheme cluster*) jako jednostki segmentacji tekstu.

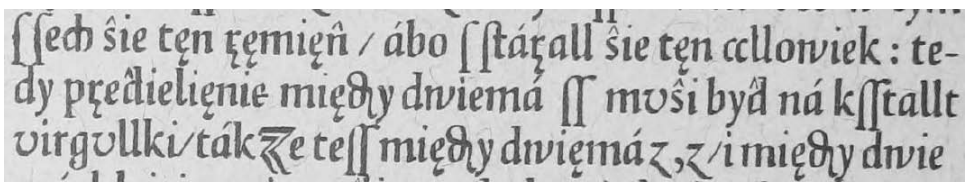
Sytuacja aktualnie wygląda więc tak, że pojęcie znaku w standardzie Unicode oddaliło się od intuicyjnego rozumienia znaku piśmiennego, a pojęcie grafemu jest zbyt abstrakcyjne. W związku z tym od pewnego czasu proponuję posługiwać się terminem *tekstel* (ang. *textel*, skrót od *text element*); analogia do terminu *piksel* (element obrazu) jest zamierzona. Dla konkretnego języka zbiór teksteli definiujemy, wskazując konkretne znaki Unicode'u lub ich sekwencje, co z kolei pozwala zdefiniować nowy sposób normalizacji tekstu: preferujemy znaki kompozytowe, a równoważność dostosowawczą stosujemy tylko do znaków niebędących tekstelami. Pojęcie to ma już konkretne praktyczne zastosowanie. Otóż przy konwersji plików poligraficznych *Słownika polszczyzny XVI wieku* na Unicode w celu udostępnienia w wyszukiwarce leksykograficznej (http://poliqarp.wbl.klf.uw.edu.pl, por. Bień 2010) standardowe metody normalizacji albo pozostawiają

[2] Nie znam, niestety, autora tego bardzo zręcznego tłumaczenia terminu *combining character*.

bez zmiany ligaturę *fi*, albo niwelują różnicę między długim i krótkim *s*. Jedno i drugie było nieakceptowalne, w celu rozpisania ligatur na składowe litery przy jednoczesnym zachowaniu długiego *s* trzeba było wprowadzić nową metodę normalizacji, uwzględniającą fakt, że długie *s* w przeciwieństwie do ligatur technicznych jest pełnoprawnym tekstelem.

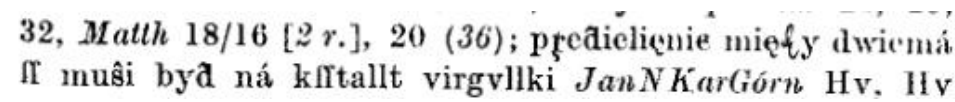
3. Unicode i polszczyzna

Załączone rysunki ilustrują pewien regres, jaki nastąpił w wyniku komputeryzacji składu *Słownika polszczyzny XVI wieku*. Na rys. 1 widzimy oryginalny fragment traktatu ortograficznego Januszowskiego (Januszowski 1594).



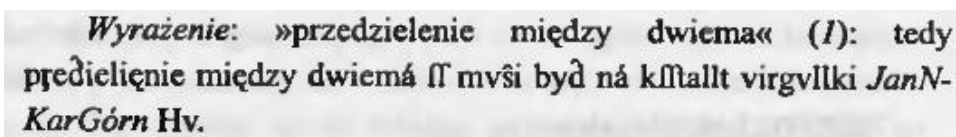
Rysunek 1. *Nowy Karakter Polski* arkusz H karta 1 verso.

W tomie opublikowanym w 1972 (rys. 2) widać próbę wiernego oddania kształtu nietypowej ligatury w słowie „między”[3].



Rysunek 2. *Słownik polszczyzny XVI wieku* tom VI, s. 233 (1972).

W tomie składanym komputerowo około 30 lat później poszerzenie repertuaru znaków najwyraźniej nie było praktycznie wykonalne, jak pokazuje to rys. 3.



Rysunek 3. *Słownik polszczyzny XVI wieku* tom XXXI, s. 341 (2003).

Pierwszym krokiem do rozwiązania takiego problemu jest rozważenie, czy mamy tu do czynienia z odmiennym kształtem jakiegoś znaku już dostępnego

[3] Ciekawe jest, jak zostało to zrealizowane technicznie — jeśli skład był wykonywany na naświetlarce Monophoto 600, to nietypowe znaki mogły być narysowane ręcznie, sfotografowane i oprawione w typowe ramki do slajdów, a następnie wykorzystywane do składu na tych samych prawach, co znaki z fabrycznego repertuaru.

w standardzie Unicode. Jeśli tak, to sprawa sprowadza się do stworzenia odpowiedniego fontu lub – lepiej – zmodyfikowania odpowiedniego fontu dostępnego na swobodnej licencji (przykładem takiego fontu może być Junicode, por. http://junicode.sourceforge.net). W razie odpowiedzi negatywnej najprościej jest przydzielić znakowi jakąś współrzędną kodową z obszaru użytku prywatnego (PUA, Private Use Area), specjalnie przeznaczonego do takich celów przez standard Unicode. Na osobisty użytek można wybrać dowolną współrzędną, ale wskazane jest skonsultowanie decyzji z innymi potencjalnymi użytkownikami. W tym wypadku właściwym forum jest MUFI, czyli Medieval Unicode Font Initiative (http://www.mufi.info), w swobodnym tłumaczeniu – inicjatywa fontów Unicode'u dla tekstów średniowiecznych.

Omawiany znak to tylko jeden z wielu dawnych polskich znaków, dla których nie ma ewidentnego sposobu reprezentowania w standardzie Unicode. Więcej przykładów można znaleźć m.in. w artykule Opalińskiego (2007) i w moim opracowaniu (Bień 2011). Aktualnie uważam, że najlepszym sposobem reprezentowania zarówno tego znaku, jak i znaku wymienionego z mojej inicjatywy na pozycji 13 w przygotowanym przez MUFI wykazie (http://www.mufi.info/pipeline), jest traktowanie ich jako odmiennego kształtu znaku U+01F3 (LATIN SMALL LETTER DZ).

4. Uwagi końcowe

Alografy i pokrewne pojęcia lingwistyki strukturalnej nie wydają się wystarczające z perspektywy uwzględnienia całej złożoności problematyki tekstów pisanych. Prawie 100 lat po wprowadzeniu pojęcia grafemu, grafemika jako nauka, moim zdaniem, ciągle jeszcze nie istnieje.

Jak wiadomo, geometria – jedna z najbardziej abstrakcyjnych dziedzin matematyki – to historycznie wiedza o mierzeniu ziemi, czyli pól uprawnych na potrzeby obliczania podatków. Nie wykluczone, że analogicznie powstanie grafemika, jako efekt rozbudowy i ewolucji aparatu pojęciowego standardu Unicode, który tutaj został przedstawiony tylko w mikroskopijnym fragmencie[4].

Literatura cytowana

Bień 1991: J. S. Bień, *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, Warszawa. Także: http://bc.klf.uw.edu.pl/12.

Bień 2010: J. S. Bień, *Dygitalizacja i komputeryzacja słowników na przykładzie Słownika polszczyzny XVI wieku* [w:] *Język polski – wczoraj, dziś, jutro*, red. B. Czopek-Kopciuch, P. Żmigrodzki, Kraków, s. 131–138. Także: http://bc.klf.uw.edu.pl/165.

[4] Niniejszy artykuł był pierwotnie przygotowany za pomocą systemu LATEX. Za pomoc przy konwersji i dostosowaniu do wymagań redakcji dziękuję Joannie Bilińskiej i Monice Kresie.

Bień 2011: J. S. Bień, *Historical Polish texts and Unicode*. *Discussion Paper:* http://bc.klf.uw.edu.pl/179.

Baudouin de Courtenay 1983: J. N. Baudouin de Courtenay, *Charakterystyka psychologiczna języka polskiego* (1915), przedruk w: *Dzieła wybrane*, t. V, PWN, Warszawa.

Character (symbol), 2011: http://en.wikipedia.org/w/index.php?oldid=404154683 (28.02.2011).

Januszowski 1594: J. Januszowski, *Nowy Karakter Polski Z Drukarnie Lazarzowey*, Kraków. Także: http://www.pbi.edu.pl/content.php?p=30709&s=1, http://www.dbc.wroc.pl/publication/4239.

Lisowski 2001: T. Lisowski, *Grafia druków polskich z 1521 i 1522 roku. Problemy wariantywności i normalizacji*, Wydawnictwo Naukowe UAM, Poznań.

Lisowski 2004: T. Lisowski, *Ideografizacja polskiego pisma a interpretacja historycznojęzykowa, czyli co wiemy o dawnym systemie graficznym*, „Biuletyn Polskiego Towarzystwa Językoznawczego" LX, s. 17–27. Także: http://www.mimuw.edu.pl/polszczyzna/PTJ/aB/b60_017-027.html.

Opaliński 2007: K. Opaliński, *Problemy kodowania korpusów historycznych (na przykładzie tekstów XVI-wiecznych)* [w:] *Z zagadnień leksykologii i leksykografii języków słowiańskich*, red. J. Kamper-Warejko, I. Kaproń-Charzyńska, Wydawnictwo UMK, Toruń, s. 107–114.

Saloni, Szpakowicz, Świdziński 1981: Z. Saloni, S. Szpakowicz, M. Świdziński, *Szkic koncepcji ogólnego słownika podstawowego współczesnej polszczyzny pisanej,* „Biuletyn Polskiego Towarzystwa Językoznawczego" XXXIX, s. 131–146.

STJ 1970: Z. Gołąb, A. Heinz, K. Polański (red.), *Słownik terminologii językoznawczej*, PWN, Warszawa.

Szpakowicz, Świdziński 1981a: S. Szpakowicz, M. Świdziński, *Formalna definicja równorzędnej grupy nominalnej we współczesnej polszczyźnie pisanej,* maszynopis powielony. Także „Studia Gramatyczne" IX (1990), s. 9–54.

Szpakowicz, Świdziński 1981b: S. Szpakowicz, M. Świdziński, *Zarys klasyfikacji schematów zdaniowych we współczesnej polszczyźnie pisanej,* „Polonica" VII, s. 5–35.

Świdziński 1979: M. Świdziński, *Klasyfikacja prostych grup syntaktycznych we współczesnej polszczyźnie pisanej.* „Studia Gramatyczne" III, s. 129–147.

Świdziński 1984: M. Świdziński, *Tzw. wypowiedzenie złożone w gramatyce formalnej współczesnej polszczyzny pisanej* [w] *Język. Teoria – dydaktyka VI*, WSP, Kielce, s. 3–41.

Świdziński 1986: M. Świdziński, *Elipsa w gramatyce formalnej współczesnej polszczyzny pisanej,* „Prace Filologiczne" XXXIII, s. 357–364.

Świdziński 1992: M. Świdziński*, Gramatyka formalna języka polskiego*, WUW, Warszawa.

Wywiałkowski 1881: Ż. Wywiałkowski, *Słowniczek wyrażeń w zawodzie czcionkarstwa polskiego używanych i używać się mogących*, Warszawa. Także: http://www.sbc.org.pl/publication/9202.

Basic elements of electronic texts
Summary

The article presents some issues concerning representation of Polish texts in a software system, with a special attention paid to Unicode standard. The paper argues for introducing the term *tekstel*, understood as referring to a generalization of the concept of sign, as understood within IT.